UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 05 - 35
Statistics and Econometrics Series 04
May 2005

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

TRANSIENT BAYESIAN INFERENCE FOR SHORT AND LONG-TAILED GI/G/1 QUEUEING SYSTEMS

María Concepción Ausín, Michael Peter Wiper and Rosa Elvira Lillo *

Abstract

In this paper, we describe how to make Bayesian inference for the transient behaviour and busy period in a single server system with general and unknown distribution for the service and interarrival time. The dense family of Coxian distributions is used for the service and arrival process to the system. This distribution model is reparametrized such that it is possible to define a non-informative prior which allows for the approximation of heavy-tailed distributions. Reversible jump Markov chain Monte Carlo methods are used to estimate the predictive distribution of the interarrival and service time. Our procedure for estimating the system measures is based in recent results for known parameters which are frequently implemented by using symbolical packages. Alternatively, we propose a simple numerical technique that can be performed for every MCMC iteration so that we can estimate interesting measures, such as the transient queue length distribution. We illustrate our approach with simulated and real queues.

Keywords: GI/G/1 queues, Bayesian inference, transient analysis, busy period, heavy tails, Reversible jump.

*Ausin, Departamento de Estadística, Universidad Carlos III de Madrid, c/ Madrid, 126, 28903 Getafe (Madrid), Spain, email: concepción.ausin@ucm.es; Wiper, Departamento de Estadística, Universidad Carlos III de Madrid, c/ Madrid, 126, 28903 Getafe (Madrid), Spain, email: michael.wiper@uc3m.es; Lillo, Departamento de Estadística, Universidad Carlos III de Madrid, c/ Madrid, 126, 28903 Getafe (Madrid), Spain, email: rosa.lillo@uc3m.es.

Transient Bayesian inference for short and long-tailed $GI/G/1$ queueing systems.

M. C. Ausín, M. P. Wiper, R. E. Lillo.

Departamento de Estadística y Econometría

Universidad Carlos III de Madrid

Madrid, 126, 28903 Getafe, Madrid, Spain

Abstract

In this paper, we describe how to make Bayesian inference for the transient behaviour and busy period in a single server system with general and unknown distribution for the service and interarrival time. The dense family of Coxian distributions is used for the service and arrival process to the system. This distribution model is reparametrized such that it is possible to define a non-informative prior which allows for the approximation of heavy-tailed distributions. Reversible jump Markov chain Monte Carlo methods are used to estimate the predictive distribution of the interarrival and service time. Our procedure for estimating the system measures is based in recent results for known parameters which are frequently implemented by using symbolical packages. Alternatively, we propose a simple numerical technique that can be performed for every MCMC iteration so that we can estimate interesting measures, such as the transient queue length distribution. We illustrate our approach with simulated and real queues.

Keywords: GI/G/1 queues, Bayesian inference, transient analysis, busy period, heavy tails, Reversible jump.

1 Introduction.

Two of the usual complicating factors in the queueing context are non-stationarity due to time-of-day effects and non-exponential service (or/and inter-arrival) times. In many practical situations, the estimation of the transient behaviour in real queues is of great interest, for example, when the system is regularly stopped and started again, or when the convergence to the steady-state is very slow. On the other hand, systems with both Poisson arrival process and exponential service times are rarely found in practice. Furthermore, it is well known that some quantities in many communication systems, such as Internet-related systems, have long-tailed distribution and cannot be represented by exponential densities. The main contributions of this work is to show how to address these difficulties which are motivated with a real data problem from an Israeli bank.

Bayesian estimation of stationary distributions in queueing systems, such as the queue size, is a fairly developed research area. Some useful references are Armero and Bayarri (1994, 1996), Ríos et al. (1998), Wiper (1998), Wiper et al. (2001), Armero and Conesa (2000, 2004) and Ausín et al. (2003, 2004) and the references therein. However, much less progress has been made on transient and busy period analysis. Some comments about the difficulties for the estimation of the transient behaviour are included in Armero and Bayarri (1998). Transient analysis of two kind of birth and death Markov processes have been recently considered in Dauxois (2004). The predictive busy period distribution for the $M/M/1$ queue is obtained in Armero and Bayarri (1994) and for the $M/G/1$ in Ausín et al. (2004) using matrix-analytic methods. Most of the queueing systems considered in the Bayesian literature assume whether Poisson arrival processes or exponential service times. To our knowledge, the only exception is Conti (2004), where discrete $Geo/G/1$ queues are analyzed.

The class of Coxian distributions is considered in this paper to describe the arrival and service processes. This distribution model is dense over the set of distributions on the positive reals, see e.g. Bertsimas (1990), and thus, any continuous and positive distribution can be approximated arbitrarily closely with a Coxian distribution by increasing the number of parameters. The Coxian family is equivalent to the mixtures of

generalized Erlang distributions (sum of exponentials with different rates) see Asmussen (2003), and thus, it includes the exponential, Erlang, hyperexponential and mixtures of Erlang distributions as particular cases. It is also equivalent to a considerable subset of the versatile class of phase-type distributions, the so-called acyclic distributions, as shown in Cumani (1982).

The Coxian distribution have been considered in a previous work, see Ausín et al. (2003). We propose a reparametrization of the Coxian mixture model following the ideas given in Robert and Mengersen (1999) for normal mixtures and in Gruet et al. (1999) for exponential mixtures. We show that using this new parametrization, it is possible to develop a non-informative approach which leads to very good approximations of long-tailed distributions, such as the Pareto or Weibull distributions. Queueing systems with heavy-tailed interarrival or service time distributions are very difficult to analyze. Abate et al. (1994) calculate the performance measures for these kind of queues using numerical Laplace transform inversion. However, it is not always posible to obtain convenient Laplace transforms of the long-tailed distributions as for the Pareto distribution. The Laplace transform of the Coxian distribution is well known and thus, many results from the queueing theory can be applied for the Bayesian estimation. Furthermore, it is known that if we approximate a given general (short or long-tailed) interarrival or service time distributions by another distribution, the performance measures such as the transient waiting time will also be approximately what it would be with the original interarrival or service time distribution, see e.g. Feldmand and Whitt (1998) in the context of exponential mixtures.

Our procedure for estimating the transient behaviour and busy period for the $GI/G/1$ queueing system is based in the results obtained by Bertsimas and Nazakato (1992) when the system parameters are known. These results involve roots of polynomial equations which can not be, in general, computed analytically and they are frequently obtained by using symbolical packages. This approach is not feasible when using reversible jump methods and not even the number of roots of the equation to solve is fixed. Alternative, we describe a simple technique to obtain the polynomial coefficients in order to numerically find the roots for every MCMC iteration so that we can estimate, for example, the Laplace transform of the busy period distribution. Numerical inversion methods are also employed to approximate the inverse Laplace transform

of the distributions under study.

This paper is organized as follows. In Section 2, we first introduce the Coxian distribution model with a new parametrization. Then, we describe how to make Bayesian inference for this model using a non-informative prior and a reversible jump MCMC algorithm, see Green (1995) and Richardson and Green (1997). In Section 3, we start briefly describing the results obtained by Bertsimas and Nazakato (1992). Then, we explain a numerical procedure to incorporate these results within the reversible jump algorithm so that we can estimate the transient and busy period distributions given the interarrival and service data. At the end of each section, our approach is illustrated with simulated and real queueing systems and some concluding remarks are also included.

2 Fitting interarrival and service time distributions.

We will assume that both interarrival and service times are independent random variables following a Coxian distribution, also called Mixed Generalized Erlang distribution (MGE), that is defined as follows. Let X be an interarrival (or service) time, then,

$$X = \begin{cases} Y_1, & \text{with probability} = P_1 \\ Y_1 + Y_2, & \text{with probability} = P_2 \\ \vdots & \vdots \\ Y_1 + ... + Y_L, & \text{with probability} = P_L \end{cases} \tag{1}$$

where $Y_r \sim \exp(\lambda_r)$, $P_r, \lambda_r > 0$ and $\sum_{r=1}^{L} P_r = 1$. This distribution model has a nice visual representation in terms of exponential phases, see Figure 1. Observe that the interarrival (or service) time of each customer can be represented by a sequence of a variable number of exponential stages.

The Coxian distribution model (1) can also be expressed as a mixture form,

$$f(x \mid L, \mathbf{P}, \boldsymbol{\lambda}) = \sum_{r=1}^{L} P_r f_r(x \mid \lambda_1, ..., \lambda_r), \qquad x \geq 0, \tag{2}$$

where $f_r(x \mid \lambda_1, ..., \lambda_r)$ is the density of a generalized Erlang distribution, i.e. the density of a sum of r

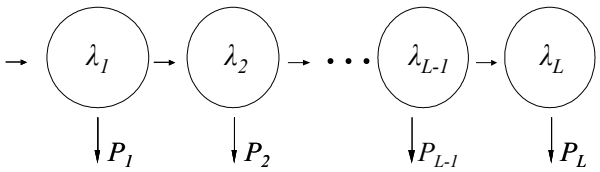

Figure 1: Graphical representation of a Coxian distribution.

exponentials with rates $\lambda_1, \lambda_2, ...$ and λ_r. If all these rates are unequal, it is given by,

$$f_r\left(x \mid \lambda_1, ..., \lambda_r\right) = \sum_{j=1}^{r} \left(\prod_{\substack{i=1 \\ i \neq j}}^{r} \frac{\lambda_i}{\lambda_i - \lambda_j} \right) \lambda_j \exp\left(-\lambda_j x\right), \qquad x \geq 0, \tag{3}$$

see Johnson and Kotz (1970). It is possible to derive alternative expressions for the general case when there are one or various groups or equal rates, but the formulas for the pdf are very complicated. In practice, when some of the rates are equal or very close to each other, it seems to be more efficient to invert numerically the Laplace transform of (3) given by,

$$f_r^*\left(s \mid \lambda_1, ..., \lambda_r\right) = \prod_{i=1}^{r} \left(\frac{\lambda_i}{\lambda_i + s} \right). \tag{4}$$

The numerical inversion can be done using the algorithm by Hosono (1981), see Appendix A, or a similar procedure. In this article, we have determined to invert the Laplace transform (4) in case of having at least a pair of rates whose difference is less than 10^{-4}.

Cumani (1982) shows that the distribution model (1) is identifiable up to permutation of the rates and then, we can assume without loss of generality that,

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_L. \tag{5}$$

The mixture model (1) have been also considered in Ausín et al. (2003) to approximate the general service time in a $M/G/c$ queue. However, in this work, the restriction (5) was not incorporated. This assumption offers much advantage to develop inference because of the identifiability of the model and because it is

possible to consider the following reparametrization,

$$\lambda_r = \lambda_1 \upsilon_2 ... \upsilon_r, \qquad \text{with } 0 < \upsilon \leq 1, \tag{6}$$

which allows improvements in the selection of the distributions a priori and in the implementation of the MCMC algorithm, as will be shown in next sections. This kind of reparametrization has also been considered in Robert and Mengersen (1999) for normal mixtures, and in Gruet et al. (1999) for exponential mixtures.

2.1 Bayesian inference.

We wish now to develop Bayesian inference considering a Coxian model for the interarrival and service time distributions. Suppose that we have observed independently n_a interarrival times and n_s service times. We assume independence between the arrival and service processes and consider independent prior distributions for the arrival and service parameters. Thus, the corresponding posterior distributions will also be independent a posteriori. This assumptions have been also considered in a number of earlier articles; see e.g. Armero and Bayarri (1996).

Therefore, we consider in this section that we have a sample of n interarrival (or service) times, $\mathbf{x} = \{x_1, ..., x_n\}$, following a Coxian distribution and we want to make inference over its parameters, $(L, \mathbf{P}, \boldsymbol{\lambda})$, that under the reparametrization (6), have been transformed into $(L, \mathbf{P}, \lambda_1, \boldsymbol{\upsilon})$. Considering first that L is known, we can assume the following improper prior distribution,

$$\begin{aligned} \mathbf{P} &\sim Dirichlet\,(1, ..., 1)\,, \\ f\,(\lambda_1, \upsilon_2, ..., \upsilon_L) &\propto \frac{1}{\lambda_1} \prod_{r=2}^{L} \upsilon_r^{-0.1}\,(1 - \upsilon_r)^{-0.1}\,. \end{aligned} \tag{7}$$

Note that the prior distribution for υ is the product of r Beta distributions, $Be\,(1.1, 1.1)$. It can be shown that this choice for the prior leads to a proper posterior distribution, see Appendix B. It is required that the parameters of the Beta distributions are larger than one for the finiteness of the predictive mean of X, as shown in Appendix B. Note that using the first parametrization, $(L, \mathbf{P}, \boldsymbol{\lambda})$, it is not possible to use an improper prior distribution where the mixture component parameters are independent from each other, see

e.g. Diebolt and Robert (1994). This type of parametrization and prior choice allows to approximate long-tailed distributions as will be shown in the examples. This is because we are not making a strong assumption about the rate of the first component, λ_1, and consequently, the means of the mixture components can be as small or as large as required.

Our task is now to construct an MCMC algorithm to sample from the joint posterior distribution. Firstly, we assume that the number of components in the mixture, k, is fixed. As is usually done in mixtures models, see e.g. Diebolt and Robert (1994), we consider a data augmentation setup introducing the component indicator variables, Z_i, such that,

$$f\left(x_i \mid Z_i = z\right) = f_r\left(x_i \mid \lambda_1, \upsilon_2, ..., \upsilon_z\right),$$

where from now on $f_r\left(x \mid \lambda_1, \upsilon_2, ..., \upsilon_z\right)$ will denote $f_r\left(x \mid \lambda_1, \lambda_2, .., \lambda_z\right)$ with the new parametrization. Then, for $i = 1, ..., n$,

$$P\left(Z_i = r \mid x_i, \mathbf{P}, \lambda_1, \boldsymbol{\upsilon}\right) \propto P_r f_r\left(x_i \mid \lambda_1, \upsilon_2, ..., \upsilon_r\right), \qquad \text{for } r = 1, ..., L, \tag{8}$$

where f_r is evaluated by using (3) or inverting (4) depending on the case. Given the missing data, $\mathbf{z} = \{z_1, ..., z_n\}$, the conditional posterior distribution of the weights is explicit and given by,

$$\mathbf{P} \mid \mathbf{x}, \mathbf{z} \sim Dirichlet(1 + n_1, ..., 1 + n_L), \tag{9}$$

where $n_r = \#\{z_i = r\}$ for $r = 1, ..., L$. The conditional posterior distributions of λ_1 and υ have not explicit expressions but their density functions can be evaluated with,

$$f\left(\lambda_1 \mid \mathbf{x}, \mathbf{z}, \boldsymbol{\upsilon}\right) \propto \prod_{i=1}^{n} f_{z_i}\left(x_i \mid \lambda_1, \upsilon_2, ..., \upsilon_{z_i}\right) f\left(\lambda_1\right), \tag{10}$$

$$f\left(\upsilon_r \mid \mathbf{x}, \mathbf{z}, \lambda_1, \boldsymbol{\upsilon}_{-r}\right) \propto \prod_{\substack{i=1 \\ z_i \geq r}}^{n} f_{z_i}\left(x_i \mid \lambda_1, \upsilon_2, ..., \upsilon_{z_i}\right) f\left(\upsilon_r\right), \qquad \text{for } r = 2, ..., L. \tag{11}$$

In Ausín et al. (2003), a larger missing data set was considered in order to obtain explicit conditional posterior distributions. However, we have observed that, in general, the performance of the algorithm is better (specially for large L) if we introduce Metropolis Hastings steps within the Gibbs sampler to generate values from (10) and (11), as shown below.

We propose the following Gibbs sampling scheme:

1. Set initial values for $\mathbf{P}^{(0)}$, $\lambda_1^{(0)}$ and $\upsilon^{(0)}$.

2. Complete the missing data sampling $\mathbf{z}^{(j)}$ from $\mathbf{z} \mid \mathbf{x}, \mathbf{P}^{(j-1)}, \lambda_1^{(j-1)}, \upsilon^{(j-1)}$.

3. Generate $\mathbf{P}^{(j)}$ from $\mathbf{P} \mid \mathbf{x}, \mathbf{z}^{(j)}$.

4. Generate $\lambda_1^{(j)}$ from $\lambda_1 \mid \mathbf{x}, \mathbf{z}^{(j)}, \upsilon^{(j-1)}$.

   Use a Metropolis step with a Gamma candidate distribution,

$$\tilde{\lambda}_1 \sim G\left(m, m/\lambda_1^{(j-1)}\right). \tag{12}$$

5. Generate $\upsilon_r^{(j)}$ from $\upsilon_1 \mid \mathbf{x}, \mathbf{z}^{(j-1)}, \upsilon_1^{(j)}, ..., \upsilon_{r-1}^{(j)}, \upsilon_{r+1}^{(j-1)}, ..., \upsilon_L^{(j-1)}$, for $r = 1, ..., L$.

   Use a Metropolis step with a Beta mixture candidate distribution,

$$g\left(\tilde{\upsilon}_r \mid \upsilon_r^{(j-1)}\right) = \frac{1}{2} Be\left(\frac{1}{1-\upsilon_r^{(j-1)}}, 2\right) + \frac{1}{2} Be\left(2, \frac{1}{\upsilon_r^{(j-1)}}\right), \qquad \text{for } r = 2, ..., L \tag{13}$$

6. $j = j + 1$. Go to 2.

In steps 4 and 5 candidates values for the conditional posterior distributions of λ_1 and υ_r are generated and accepted with probability,

$$\alpha = \min\left\{\frac{f\left(\tilde{\theta} \mid \cdot\right)}{f\left(\theta^{(j-1)} \mid \cdot\right)} \frac{g\left(\theta^{(j-1)} \mid \tilde{\theta}_r\right)}{g\left(\tilde{\theta} \mid \theta_r^{(j-1)}\right)}\right\},$$

where θ represents one of the parameters λ_1 or υ_r; f is given in (10) and (11), respectively, and g is its candidate distribution given in (12) and (13), respectively. In step 4, the choice for (12) is based on the similarity of (10) with a Gamma distribution. Note that λ_1 follows a Gamma posterior distribution when a larger missing data set is considered, see Ausín et al. (2003). The value of m can be chosen to give an adequate acceptance rate. We have found in practice that $m = 2$ seems to be appropriate. In step5, the beta mixture (13) has been chosen to avoid generating indefinitely values for υ_r near to zero or one and simultaneously preserve the mode of the value of υ_r in the previous iteration, see Wiper (?).

We can extend the previous Gibbs sampling algorithm to the case where k is unknown. First, we assume a discrete uniform prior defined on $[1, k_{\max}]$. Other choices such as Poisson or geometric prior distributions are useful when low values of k are desired to be favoured a priori. In order to let the chain move through the posterior distribution of k, we make use of the reversible jump technique introduced by Green (1995) and adapted for normal mixtures in Richardson and Green (1997). Specific moves in the parameters are needed to be defined to allow changing the number of components from k to $k \pm 1$. We consider the so called split and combine moves where one mixture component, r, is splitted into two adjacent components (r_1, r_2) or two adjacent components are combined into one, respectively. In the combine move the parameters are modified such that,

$$\tilde{P}_r = P_{r_1} + P_{r_2}, \qquad \tilde{\upsilon}_r = \upsilon_{r_1}\upsilon_{r_2},$$

which implies that $\tilde{\lambda}_r = \lambda_{r_2}$. For the case that $r = 1$ we consider $\tilde{\lambda}_1 = \lambda_1\upsilon_2$. For the split move,

$$\tilde{P}_{r_1} = u_1 P_r, \qquad \tilde{P}_{r_2} = (1 - u_1)P_r,$$

where $u_1 \sim U(0, 1)$ and,

$$\tilde{\upsilon}_{r_1} = u_2 + \upsilon_r (1 - u_2), \qquad \tilde{\upsilon}_{r_2} = \frac{\upsilon_r}{u_2 + \upsilon_r (1 - u_2)},$$

where $u_2 \sim U(0, 1)$, which implies that,

$$\tilde{\lambda}_{r_1} = \lambda_{r-1}u_2 + \lambda_r (1 - u_2), \qquad \tilde{\lambda}_{r_2} = \lambda_r.$$

For the case that $r = 1$, we consider $\tilde{\lambda}_1 = \lambda_1/u_2$ and $\tilde{\upsilon}_2 = u_2$ where $u_2 \sim U(0.5, 1)$. Also, every observation such that $z_i = r$ is assigned to any of the two components, r_1 or r_2, with probability,

$$P\left(\tilde{Z}_i = r_j\right) \propto \tilde{P}_{r_1} f_{r_1}\left(x_i \mid \lambda_1, \upsilon_2, ..., \tilde{\upsilon}_{r_j}\right), \qquad \text{for } j = 1, 2.$$

Note that the defined split-combine moves are chosen such that the parameters in the remaining mixture components are not modified, as considered in Gruet et al. (1999), and do not necessarily preserve the moments of the distribution of X. The acceptance probability of a split move is $\min\{1, A\}$ where,

$$A = \frac{\tilde{P}_{r_1}^{\tilde{n}_{r_1}} \tilde{P}_{r_2}^{\tilde{n}_{r_2}}}{P_r^{n_r}} \frac{\prod_{i:\tilde{z}_i \geq r_1} f_{\tilde{z}_i}\left(x_i \mid \lambda_1, \upsilon_2, ..., \tilde{\upsilon}_{\tilde{z}_i}\right)}{\prod_{i:z_i \geq r_1} f_{z_i}\left(x_i \mid \lambda_1, \upsilon_2, ..., \upsilon_{z_i}\right)} \times \frac{d_{L+1}}{b_L \prod_{i:z_i = r} P\left(\tilde{Z}_i = \tilde{z}_i\right)} \times \frac{P_r (1 - \upsilon_r)}{u_2 + \upsilon_r (1 - u_2)}$$

when $r > 1$ and where $\tilde{n}_{r_j}$ is the number of observations assigned to component r_j, for $j = 1, 2$ and d_L and b_L are respectively the probabilities of a combine or a split move. The last factor is the determinant of the Jacobian of the transform from $(P_r, \upsilon_r, u_1, u_2)$ to $(\tilde{P}_{r_1}, \tilde{P}_{r_2}, \tilde{\upsilon}_{r_1}, \tilde{\upsilon}_{r_2})$. The acceptance probability for the reverse combine move can be obtained analogously. Note that the acceptance probabilities does not need to incorporate factorial terms due to the natural order of the rates derived from the new parametrization, see Gruet et al. (1999). As usual, given a MCMC sample of size J, the predictive distribution of the interarrival (or service) time can be approximated by,

$$f(x \mid \mathbf{x}) \approx \frac{1}{J} \sum_{j=1}^{J} \sum_{r=1}^{L^{(j)}} P_r^{(j)} f_r \left(x \mid \lambda_1^{(j)}, \upsilon_2^{(j)}, ..., \upsilon_r^{(j)} \right). \tag{14}$$

2.2 Results for simulated and real data sets.

In this section, we illustrate the performance of the proposed Bayesian density estimation method using different data samples. We consider four simulated and four real data sets. For the simulated case, we generate two samples of short-tailed and two of long-tailed distributions. For the real case, we make use of some interarrival and service data from a face-to-face bank data base downloaded from the Professor Mandelbaum's web page, `http://iew3.technion.ac.il/serveng`.

2.2.1 Simulated examples.

We generate 300 data for each of the following distributions:

1. A single exponential distribution with $\lambda = 1$.

2. A Coxian distribution with $\mathbf{P} = (0.09, 0.7, 0.01, 0.2)$ and $\boldsymbol{\lambda} = (1.1, 1.0, 0.251, 0.25)$.

3. A Weibull distribution, see (15), with $c = 0.3$ and $a = 9.26053$.

4. A Pareto distribution, see (16), with $a = 2.2$ and $b = 0.8333$.

The exponential distribution is the simplest Coxian case. In case 2, we have chosen very close rates to illustrate that this does not affect to the stability of the algorithm as commented in Section 2. These

two first examples are short-tailed distributions. Note that the Coxian distribution is short-tailed as any phase-type distribution is. However, we show how the Coxian model can be used to approximate a long-tail behaviour using eventually a large number of components, L. Two well-known examples of heavy-tailed distributions are the Pareto and the Weibull (with scale parameter, c, less than one) distributions. The Weibull cumulative distribution function is given by,

$$F(x) = 1 - \exp\{-(ax)^c\} \qquad x > 0, \tag{15}$$

and the Pareto cumulative distribution function is,

$$F(x) = 1 - \frac{1}{(1+bx)^\alpha}, \qquad x > 0. \tag{16}$$

The cases 3 and 4 are two of the examples considered in Feldman and Whitt (1998) where the parameters are chosen such that their means are equal to one.

Figure 2 illustrates the empirical against the estimated cdf after running the MCMC algorithm for 100000 burn-in iterations followed by an additional 100000 iterations. Note that the abscissae axis in cases 3 and 4 are displayed in log scale. We can observe that the fits are quite satisfactory even for the heavy-tailed cases. The proportions of moves accepted vary from 7 to 15 per cent which are reasonable values in reversible jump setups, see Richardson and Green (1997).

Figure 3 shows the posterior probabilities of the number of components, L. Observe that in cases 1 and 2, the algorithm identifies the correct mixture size and the posterior mode of L is equal to its true value. The estimated density in case 3 requires a large number of phases to fit the Weibull distribution. However, it is smaller than the 20 mixture components used in Feldman and Whitt (1998) to fit a hyperexponential distribution. This benefit is better illustrated in case 4 where the posterior mode of L is 3 in contrast with the 14 exponential components used in Feldman and Whitt (1998). We have also compared cases 3 and 4 with the Bayesian algorithm for exponential mixtures proposed in Gruet et al. (1999) and the posterior mode of L results to be one unit larger than using the Coxian distribution. In addition, note that, unlike the exponential mixtures, the Coxian model can also fit densities with non zero mode such as the given in case 3 and in a real example below.

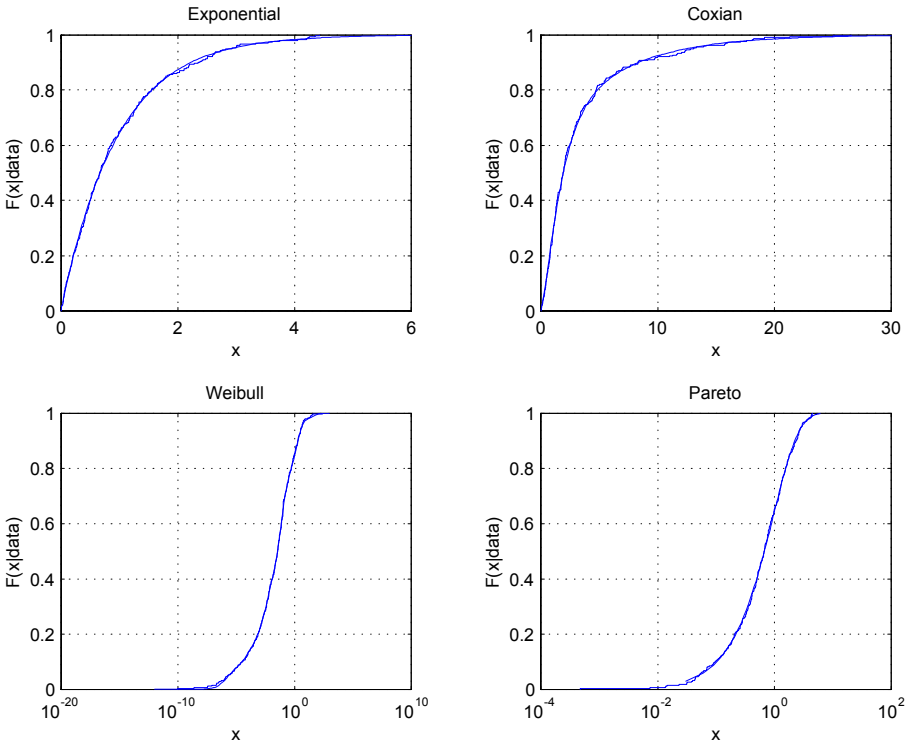

Figure 2: Empirical cummulative distribution functions in comparison with the predictive cumulative distributions. The Weibull and Pareto distributions are displayed in log scale.

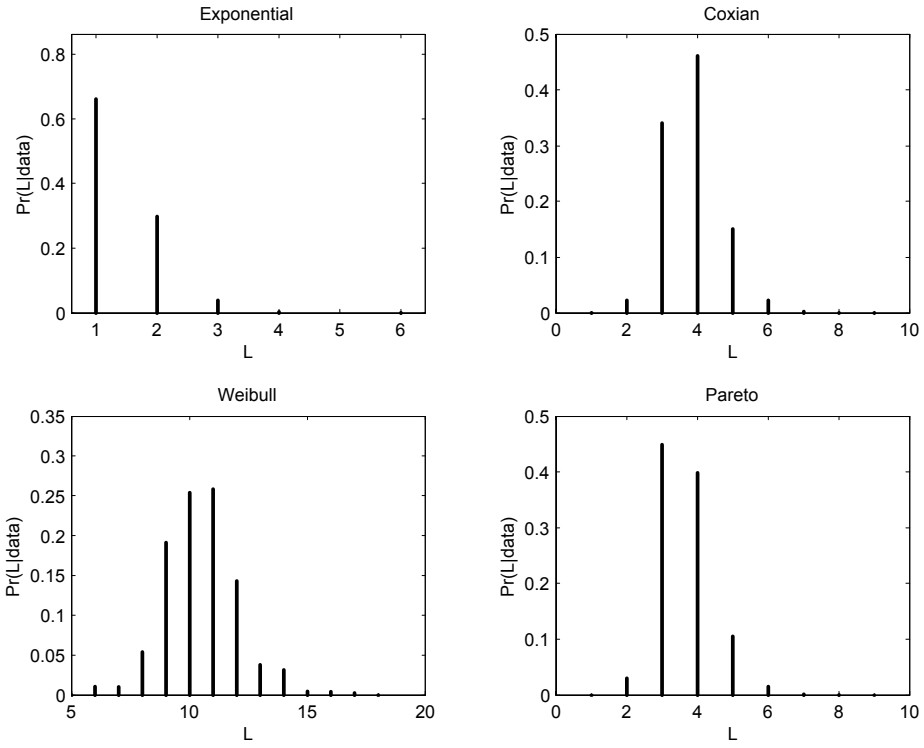

Figure 3: Posterior probabilities of the mixture size, L.

2.2.2 Real data application.

We now illustrate the method with real interarrival and service times taken from a branch of an Israeli bank. We consider data of two different kind of services: foreign currency exchange and business banking transactions. This type of services require a single server that works three days a week from 8:30 to 12:00 and two days a week from 8:30 to 12:30 and from 16:00 to 18:00. The data are recorded during 14 days and consist of 249 interarrival and 270 service times for foreign currency exchange and 822 interarrival and 843 service times of business banking transactions.

Figure 4 shows the histograms of the observed data and the predictive densities obtained after a run of 100000 iterations in equilibrium. Observe that the Coxian distribution can also fit properly non-monotone densities as the business banking service time density. Interarrival times between customers asking for currency exchange are larger on average than for business transactions. Their predictive means are 12.3 and 4.70 minutes, respectively. Also the required time for exchange currency services is greater on average than for business transaction services. Their predictive means are 7.22 and 3.86 minutes, respectively. However, in this case the business banking service distribution seems quite heterogeneous, with a maximum value of 42.283 minutes, while the currency exchange services seems fairly homogeneous.

Table 3 gives the posterior distribution of L for the four data sets. Observe that our Bayesian density estimation method predicts with some uncertainty an exponential distribution for the foreign currency service time and for the business banking interarrival time with estimated rates approximately equal to 0.14 and 0.22, respectively. For the exchange currency interarrival distribution, the algorithm suggests a two component mixture . The first component is an exponential and the second an Erlang distribution both with rates close to 0.1. Finally, the business banking service time distribution requires a fairly large number of components to be fitted.

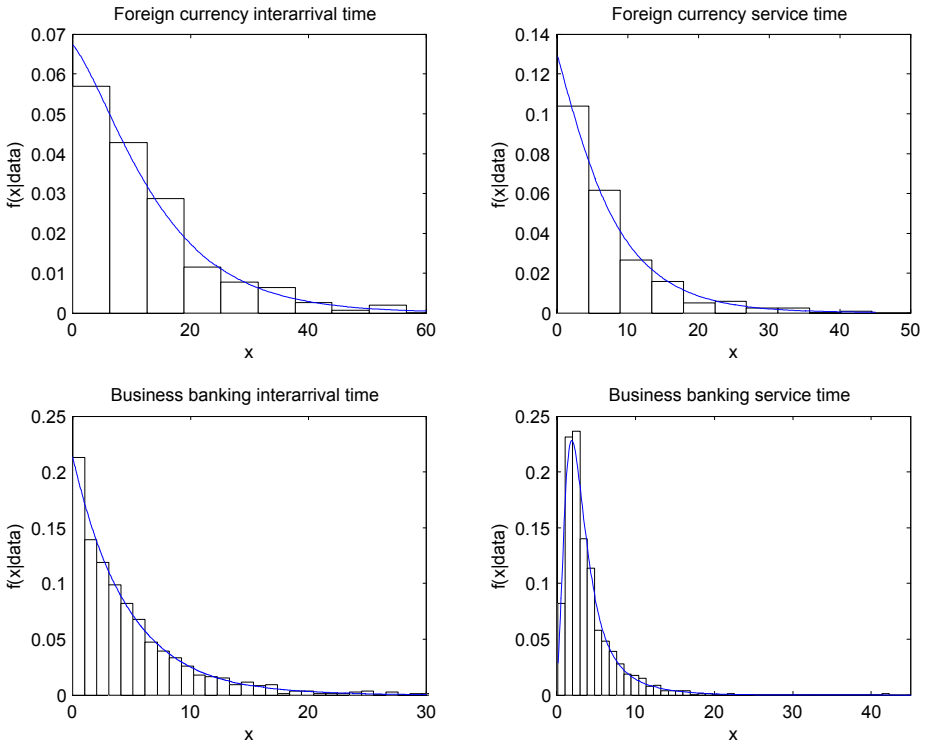

Figure 4: Predictive interarrival and service time densities for the four real data sets.

	Foreign Currency		Business Banking	
$P(L \mid data)$	Arrival	Service	Arrival	Service
1	.22331	.50194	.69997	.00000
2	.64507	.37011	.26687	.00000
3	.11807	.10479	.03198	.00000
4	.01151	.02124	.00108	.33684
5	.00169	.00186	.00010	.42807
6	.00026	.00006	.00000	.18957
7	.00009	.00000	.00000	.04314
8	.00000	.00000	.00000	.00238
9	.00000	.00000	.00000	.00000

Table 1: Posterior probabilitities of the number of components, L, for the four real data sets.

3 Bayesian prediction for the $GI/G/1$ queueing model.

In this section, we are interested in the performance of the $GI/G/1$ queueing model. Therefore, we will assume that the interarrival time, A_1, and the service time, S_1, are distributed as Coxian distributions with parameters $\boldsymbol{\theta}_\lambda = \{L, \mathbf{P}, \boldsymbol{\lambda}\}$ and $\boldsymbol{\theta}_\mu = \{M, \mathbf{Q}, \boldsymbol{\mu}\}$, respectively. Figure 5 illustrates the behaviour of this queueing system. Observe that each customer must go through $1, 2, ...$ or L exponential stages of the arrival timing channel (ATC) with probabilities $P_1, P_2, ...$ or P_L, respectively, before accessing to the waiting line or, eventually, to the service timing channel (STC), where each service time is the sum of $1, 2, ...$ or M exponential stages with probabilities $Q_1, Q_2, ...$ or Q_M, respectively.

Assume now that we are given two sets of interarrival and service times, $\mathbf{t} = \{t_1, ..., t_{n_a}\}$ and $\mathbf{s} = \{s_1, ..., s_{n_s}\}$, respectively. Then, using the MCMC output obtained from the algorithm described in the previous section, we can estimate some measures of interest in the queue. An important measure of the average occupancy is

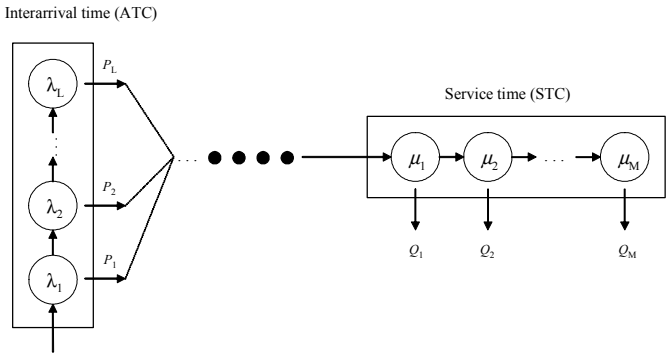

Figure 5: Ilustration of the MGE/MGE/1 queueing model.

the traffic intensity, ρ, that for this queueing model is given by,

$$\rho = \frac{E[S_1]}{E[A_1]} = \frac{\Sigma_{r=1}^{M}\left(1-\Sigma_{s=1}^{r-1}Q_s\right)\frac{1}{\mu_r}}{\Sigma_{r=1}^{L}\left(1-\Sigma_{s=1}^{r-1}P_s\right)\frac{1}{\lambda_r}}, \tag{17}$$

and it is well known that the queue is stable if $\rho < 1$, see e.g. Gross and Harris (1985). The posterior probability of having a stable queue can be estimated with,

$$P(\rho < 1 \mid \mathbf{t}, \mathbf{s}) \approx \frac{1}{J}\#\left\{\rho^{(j)} < 1\right\}, \tag{18}$$

where $\rho^{(j)}$ is the value of (17) for each element of the MCMC sample. Usually, if this probability is large enough, it is assumed that the system is stable, see e.g. Ausín et al. (2004). However, even in a stable system, it is recommended to make inference on the transient behaviour because the convergence to the steady state can be very slow or because there can be exogenous changes such that the stationary distributions do not give a realistic description of the queueing performance. The posterior mean of ρ can be estimated analogously with,

$$E[\rho \mid \mathbf{t}, \mathbf{s}] \approx \frac{1}{J}\sum_{j=1}^{J}\rho^{(j)}. \tag{19}$$

Note that this expectation is finite because the means of the predictive interarrival and service time distributions are finite as shown in Appendix B. Analogously, we can estimate the posterior mean of ρ assuming equilibrium, $E[\rho \mid \rho < 1, \mathbf{t}, \mathbf{s}]$, by simply rejecting the draws larger than one.

Given the arrival and service data, we will show in this section how to estimate the transient system size

and waiting time distributions and the length of the busy period distribution. Firstly, we will introduce some notation and results obtained by Bertsimas and Nazakato (1992) which consist of the Laplace transforms of these distributions when the system parameters are known. Then, we describe a numerical technique to extract the roots of some polynomial equations involved in these Laplace transforms. Finally, we explain how to combine this numerical procedure with the reversible jump methodology and with Laplace transform inversion methods in order to estimate the distributions of the quantities of interest.

3.1 Preliminaries.

Let A_k be the random variable representing the remaining time a customer requires to access to the waiting line if case of being in the kth stage of the ATC. Note that A_1 is the whole interarrival time. Its Laplace transform is given by,

$$f_{A_k}^*(s) = \int_0^\infty e^{-st} f_{A_k}(t)\, dt = \sum_{r=k}^{L} \frac{P_r}{1-\Sigma_{s=1}^{k-1} P_s} \prod_{i=k}^{r} \left(\frac{\lambda_i}{\lambda_i + s}\right), \qquad \text{for } k = 1, ..., L. \tag{20}$$

An analogous random variable, S_k, and its Laplace transform, $f_{S_k}^*(s)$, can be considered for the STC .

Let $A_{k,r}(x)$ be the probability that a customer in the ATC move from stage k to the stage r during the interval t without any new arrival. Note that this probability is zero if $k > r$. Its Laplace transform is given by,

$$A_{k,r}^*(s) = \int_0^\infty e^{-st} A_{k,r}(t)\, dx = \frac{1-\Sigma_{s=1}^{r-1} P_s}{1-\Sigma_{s=1}^{k-1} P_s} \frac{\prod_{i=k}^{r-1} \lambda_i}{\prod_{i=k}^{r} (\lambda_i + s)}, \qquad \text{for } r = k, ..., L. \tag{21}$$

An analogous probability, $S_{k,r}(x)$, and its Laplace transform, $S_{k,r}^*(s)$, can be considered for the STC.

Bertsimas and Nazakato (1992) formulate the queueing system as a continuous time Markov chain with infinite state space,

$$\{(N(\tau), R_a(\tau), R_s(\tau)), \quad N(\tau) = 0, 1, ...; \quad R_a(\tau) = 1, .., L; \quad R_s(\tau) = 1, ..., M.\},$$

where $N(\tau)$ denotes the number of customers in the system at time τ, $R_a(\tau)$ the ATC stage currently occupied by the arriving customer at time τ and $R_s(\tau)$ is the STC stage who is being served at time τ. Let us denote,

$$\pi_{n,i,j}(\tau) = \Pr(N(\tau) = n,\ R_a(\tau) = i,\ R_s(\tau) = j), \qquad \text{if } n \geq 1,$$

and

$$\pi_{0,i}(\tau) = \Pr(N(\tau) = 0,\ R_a(\tau) = i), \qquad \text{if } n = 0. \tag{22}$$

Assuming that the system is initially empty and $\rho < 1$, Bertsimas and Nazakato (1992) show that the Laplace transforms of these probabilities are given by,

$$\pi^*_{n,i,j}(s) = \int_0^\infty e^{-s\tau}\pi_{n,i,j}(\tau)\,d\tau = \sum_{r=1}^{M} D_r S^*_{1,j}(x_r(s))\,A^*_{1,i}(s - x_r(s))\left(f^*_{A_1}(s - x_r(s))\right)^{n-1}, \tag{23}$$

for $i = 1, ..., L$, and $j = 1, ..., M$ and,

$$\pi^*_{0,i}(s) = \int_0^\infty e^{-s\tau}\pi_{0,i}(\tau)\,d\tau = \sum_{k=1}^{L}\pi_{0,k}(0)\,A^*_{k,i}(x_r(s)) + \sum_{r=1}^{M}\frac{D_r f^*_{S_1}(x_r(s))}{x_r(s)}\left(A^*_{1,i}(s - x_r(s)) - A^*_{1,i}(s)\right), \tag{24}$$

where,

$$D_r = \frac{\sum_{k=1}^{L}\pi_{0,k}(0)\,f^*_{A_k}(x_r(s))}{1 - f^*_{A_1}(s)}\frac{(-1)^M S^*_{1,M}(0)}{S^*_{1,M}(x_r(s))}x_r(s)\prod_{\substack{k=1\\k\neq r}}^{M}\frac{x_r(s)}{x_r(s) - x_k(s)} \tag{25}$$

and $x_r(s) \equiv x$, with $r = 1, ..., M$, are the M roots of the equation,

$$\left\{\begin{array}{c} f^*_{A_1}(s - x)\,f^*_{S_1}(x) = 1, \\ \\ \mathrm{Re}(x) < 0 \text{ for } \mathrm{Re}(s) > 0. \end{array}\right\} \tag{26}$$

On the other hand, assuming the same conditions and that the elements of the initial probability vector are given by,

$$\pi_{0,i}(0) = \frac{\left(1 - \Sigma_{s=1}^{i-1}P_s\right)\frac{1}{\lambda_i}}{\Sigma_{r=1}^{L}\left(1 - \Sigma_{s=1}^{r-1}P_s\right)\frac{1}{\lambda_r}}, \qquad \text{for } i = 1, ..., L, \tag{27}$$

Bertsimas and Nazakato (1992) show that the Laplace transform of the waiting time, $W(\tau)$, of a customer arriving at time τ is given by,

$$\int_0^\infty e^{-s\tau}\Pr(W(\tau) \leq w)\,d\tau = \frac{1}{s} + \sum_{r=1}^{M}\frac{(-1)^M}{s}\frac{S_{1,M}(0)}{S_{1,M}(x_r(s))}\left(\prod_{\substack{k=1\\k\neq r}}^{M}\frac{x_r(s)}{x_r(s) - x_k(s)}\right)e^{x_r(s)w}, \tag{28}$$

where $x_r(s)$, for $r = 1, ..., M$, are the M roots of equation (26). Bertsimas and Nazakato (1992) also show that the condition (27) implies that the arrival process is in the steady state at time $\tau = 0$. Note that this condition simplifies the expression for the coefficients given in (25) as follows,

$$D_r = \frac{1}{sE[A_1]}\frac{(-1)^M S^*_{1,M}(0)}{S^*_{1,M}(x_r(s))}x_r(s)\prod_{\substack{k=1\\k\neq r}}^{M}\frac{x_r(s)}{x_r(s) - x_k(s)}, \qquad \text{for } r = 1, ..., M,$$

see Bertsimas and Nazakato (1992), and then, the probabilities given in (27) will be also assumed to obtain the distribution of $N(\tau)$ in the examples.

Finally, Bertsimas and Nazakato (1992) show that the Laplace transform of the distribution function of the length of the busy period is given by,

$$F_B^*(s) = \int_0^\infty e^{-st} \Pr(B \leq t)\, dt = \frac{f_B^*(s)}{s} = \frac{1}{s} - \frac{1 - f_{S_1}^*(s)}{s} \frac{\prod_{k=1}^M (s + \mu_k)}{\prod_{r=1}^M (s - x_r(s))} \tag{29}$$

where, again, $x_r(s)$, for $r = 1, ..., M$, are the M roots of equation (26).

3.2 Numerical extraction of the roots of the equation (26).

In this section, we describe a procedure for numerically extracting the M roots of the equation (26) in order to evaluate the Laplace transforms of $N(\tau)$, $W(\tau)$ and B in each MCMC iteration. Given a set of system parameters, $\boldsymbol{\theta} = \{L, \mathbf{P}, \boldsymbol{\lambda}, M, \mathbf{Q}, \boldsymbol{\mu}\}$, we want to solve the equation (26) that, considering (20), can be expressed as,

$$\sum_{r=1}^{L} P_r \prod_{i=1}^{r} \left(\frac{\lambda_i}{\lambda_i + s - x} \right) \times \sum_{t=1}^{M} Q_t \prod_{j=1}^{t} \left(\frac{\mu_j}{\mu_j + s - x} \right) = 1, \tag{30}$$

such that $\mathrm{Re}(x) < 0$. We could make use of a numerical algorithm such as Newton-Raphson method to solve this equation. However, note that (30) is a polynomial equation whose roots are the $L + M$ roots of the following polynomial where M of these roots verify that $\mathrm{Re}(x) < 0$,

$$\mathcal{P}(x) = \left[\sum_{r=1}^{L} \sum_{t=1}^{M} P_r Q_t \left(\prod_{i=1}^{r} \lambda_i \prod_{j=1}^{t} \mu_j \right) \mathcal{Q}_{r,t}(x) \right] - \mathcal{Q}_{0,0}(x), \tag{31}$$

where $\mathcal{Q}_{r,t}(x)$ is another polynomial given by,

$$\begin{aligned} \mathcal{Q}_{r,t}(x) &= \prod_{i=r+1}^{L} (\lambda_i + s - x) \prod_{j=t+1}^{M} (\mu_j + x) \\ &= (-1)^{L-r} \prod_{i=r+1}^{L} (x - \lambda_i - s) \prod_{j=t+1}^{M} (x + \mu_j). \end{aligned} \tag{32}$$

Therefore, we have considered the Laguerre method, see e.g. Ralston and Rabinowitz (1978), which is an algorithm design specifically to find the roots of a complex polynomial given its coefficients. In order to calculate the coefficients of the polynomial (31), we consider the Taylor expansion of $\mathcal{P}(x)$ in $x = 0$, such

that each polynomial coefficient is given by $\mathcal{P}^{(n)}(0)/n!$, for $n = 0, ..., L+M$, where,

$$\mathcal{P}^{(n)}(0) = \left[\sum_{r=1}^{L}\sum_{t=1}^{M} P_r Q_t \left(\prod_{i=1}^{r}\lambda_i \prod_{j=1}^{t}\mu_j\right)\mathcal{Q}_{r,t}^{(n)}(0)\right] - \mathcal{Q}_{0,0}^{(n)}(0). \tag{33}$$

Thus, we only need to be able to compute the nth derivatives of the polynomial $\mathcal{Q}_{r,t}(x)$ given in (32). Note that this is a polynomial of order $(L+M-r-t)$ whose roots are given by $\{(\lambda_{r+1}+s), ..., (\lambda_L+s), -\mu_{t+1}, ..., -\mu_M\}$ and the coefficient of order $(L+M-r-t)$ is $(-1)^{L-r}$. We show below how to compute the nth derivative of a polynomial given its roots and the largest order coefficient so that it is possible to implement a routine for evaluating $\mathcal{Q}_{r,t}^{(n)}(0)$ in (33).

Let $\mathcal{Q}(x)$ a polynomial of order N,

$$\mathcal{Q}(x) = a_N x^N + a_{N-1}x^{N-1} + ... + a_1 x + a_0,$$

whose N roots are $\{x_1, ..., x_N\}$, such that,

$$\mathcal{Q}(x) = a_N (x - x_1) ... (x - x_N). \tag{34}$$

The Viètas formulas, see e.g. Borwein (1995), allows to obtain the coefficients of $\mathcal{Q}(x)$ which are given by,

$$a_{N-k} = (-1)^k a_N \sum_{(i_1,...,i_k)\in\mathcal{C}_{n,k}} \left(\prod_{i=i_1}^{i_k} x_i\right), \qquad \text{for } k = 1, ..., N,$$

where $\mathcal{C}_{n,k}$ is the set of the $\binom{n}{k}$ combinations of the n elements, $\{x_1, ..., x_n\}$, taking k at a time. Observe that, given a_N, the remaining coefficients can be derived recursively with:

1. `Set` $a_{N-k} = 0$, `for` $k = 1, ..., N$.

2. `For` $k = 1, ..., N$,
$$\begin{pmatrix} a_{N-1} \\ \vdots \\ a_{N-k} \end{pmatrix} = \begin{pmatrix} a_{N-1} \\ \vdots \\ a_{N-k} \end{pmatrix} + x_k \begin{pmatrix} a_{N-1} \\ \vdots \\ a_{N-k} \end{pmatrix}.$$

Finally, once we know the coefficients, it is straightforward to see that the nth derivatives of the polynomial $\mathcal{Q}(x)$ are given by,

$$\mathcal{Q}^{(n)}(x) = \sum_{i=n}^{N} a_{N+n-i}\left(\prod_{j=1}^{n}(N-i+j)\right)x^{N-i},$$

and then,

$$\mathcal{Q}^{(n)}(0) = \sum_{i=n}^{N} a_{N+n-i} \left(\prod_{j=1}^{n} (N-i+j) \right).$$

3.3 Estimation of the transient behaviour and the busy period.

Given a sample realization of the posterior distribution of $\boldsymbol{\theta} = \{L, \mathbf{P}, \boldsymbol{\lambda}, M, \mathbf{Q}, \boldsymbol{\mu}\}$, the natural way of estimating the predictive distributions is using Monte Carlo approximations. For example, we can estimate the transient distribution of the system size, $N(\tau)$, with,

$$\Pr(N(\tau) = n \mid \mathbf{t}, \mathbf{s}, \rho < 1) \approx \frac{1}{R} \sum_{j:\rho^{(j)}<1} \Pr\left(N(\tau) = n \mid \boldsymbol{\theta}^{(j)}\right), \tag{35}$$

where $\Pr\left(N(\tau) = n \mid \boldsymbol{\theta}^{(j)}\right)$ can be obtained using a numerical inversion method such as the algorithm by Hosono (1981), see Appendix A, in order to numerically invert its Laplace transform which is given by,

$$\boldsymbol{\pi}_n^*\left(s \mid \boldsymbol{\theta}^{(j)}\right) = \begin{cases} \sum_{i=1}^{L} \pi_{0,i}^*(s), & \text{if } n = 0, \\ \sum_{i=1}^{L} \sum_{j=1}^{M} \pi_{n,i,j}^*(s), & \text{if } n \geq 1, \end{cases} \tag{36}$$

where $\pi_{0,i}^*(s)$ and $\pi_{n,i,j}^*(s)$ are given in (24) and (23), respectively, and $R = \#\{\rho^{(j)} < 1\}$. Observe that we have assumed that $\rho < 1$ in (35) because it is a required condition to apply the results obtained in Bertimas and Nazakato (1992).

Analogously, we can estimate the transient distribution function of the waiting time in the queue by,

$$\Pr(W(\tau) \leq w \mid \mathbf{t}, \mathbf{s}, \rho < 1) \approx \frac{1}{R} \sum_{j:\rho^{(j)}<1} \Pr\left(W(\tau) \leq w \mid \boldsymbol{\theta}^{(j)}\right), \tag{37}$$

where $\Pr\left(W(\tau) \leq w \mid \boldsymbol{\theta}^{(j)}\right)$ is obtained inverting numerically its Laplace transform given in (28) and the predictive distribution of the busy period,

$$F_B(x \mid \mathbf{t}, \mathbf{s}, \rho < 1) \approx \frac{1}{R} \sum_{j:\rho^{(j)}<1} F_B\left(x \mid \boldsymbol{\theta}^{(j)}\right), \tag{38}$$

where $F_B\left(x \mid \boldsymbol{\theta}^{(j)}\right)$ is obtained inverting numerically its Laplace transform given in (29).

It can be shown that the estimation of the moments of the stationary system size and waiting time distributions and the moments of the busy period distribution do not exist with the prior structure that we

have considered, see e.g. Wiper (1998). Then, the moments of the transient distributions will converge to infinity as τ goes to infinity. However, we can always estimate the median and quantiles of these distributions.

3.4 Results for simulated and real queues.

In this Section, we illustrate the behaviour of the proposed method with several simulated and real queues. Most of the simulated interarrival and service times were introduced in Section 2. Also, the two real queueing systems in the Israeli bank are analyzed.

3.4.1 Simulated queues.

We consider interarrival and service data simulated from the following two queueing systems:

- An $M/M/1$ system where the service times are the exponential data simulated in Section 2.
- A $Coxian/Pareto/1$ system where both interarrival and service times were simulated in Section 2.

For the interarrival time in the $M/M/1$ queue, we simulate 300 exponential data with mean 3.456, which is approximately equal to the mean of the Coxian interarrival time. As both systems have also the same service mean time, equal to one, they have also the same traffic intensity, $\rho = 0.289$. Table 2 shows the probabilities of having equilibrium in the system and the posterior means of ρ for both queues, see (18) and (19). Also shown are the MLE estimations of ρ. Observe that the estimations are very close to the true value.

	$P(\rho < 1 \mid data)$	$E[\rho \mid data]$	$E[\rho < 1 \mid \rho < 1, data]$	$\hat{\rho}_{MLE}$
$M/M/1$	.99998	.29009	.29008	.28935
$Coxian/Pareto/1$	.99993	.30740	.30736	.28116

Table 2: Posterior probabilitities that the system is stable and posterior mean values for the traffic intensity for the two simulated systems.

Figure 6 illustrates in dotted lines the estimated transient distributions of the queue length and waiting time, see (35) and (37), as a function of time, τ, for both simulated queues. Note that the estimations clearly converge to their stationary distributions as τ goes to infinity. Observe that, as commented before, the convergence can be slow even if the probability of equilibrium is very large. This is specially the case of the waiting time in the $Coxian/Pareto/1$ queue. Also note that although the traffic intensity is the same for both systems, the convergence to the stationary waiting time is much slower in the second than in the first queue. On the contrary the speed of convergence of the transient queue length distribution is similar in both examples. These differences are mostly originated by the long-tailed behaviour of the Pareto service time distribution. It is well known that, in general, the waiting time distributions are very influenced by the shape of the service time densities, while the queue length distributions are strongly dependent only on their first moment. Note that as τ goes to infinity, the estimated probability that the queue length is 0 approaches to one minus the posterior mean of ρ, which is coherent with the known result, $P(N=0) = 1-\rho$, where N denotes the equilibrium queue length in a GI/G/1 model, see e.g. Gross and Harris (1985). Also, the waiting time probability, $P(W(\tau) > 0)$ approaches to the estimated traffic intensity but only in the M/M/1 queue because, as it is well known, the result $P(W=0) = 1-\rho$ is only true for Poisson arrivals.

Figure 7 shows the estimated distribution functions of the busy period in dotted lines, see (38). For the $M/M/1$ case, it is compared with the theoretical distribution given the parameters in solid line. This is not done for the second case as the theoretical busy period distribution has not been obtained so far, see e.g. Gross and Harris (1885). Note that the tail of the busy period distribution is shorter for the Markovian queue.

3.4.2 Real queues.

We now analyze the real data about foreign currency exchange and business banking transactions in the Israeli bank. As the bank has a single teller for each kind of service, we have two single server, FIFO, queueing systems. For both systems, the estimated posterior probability that the system reach the steady-state is extremely high and the posterior mean values for ρ are close to the MLE estimators, see (18) and (19),

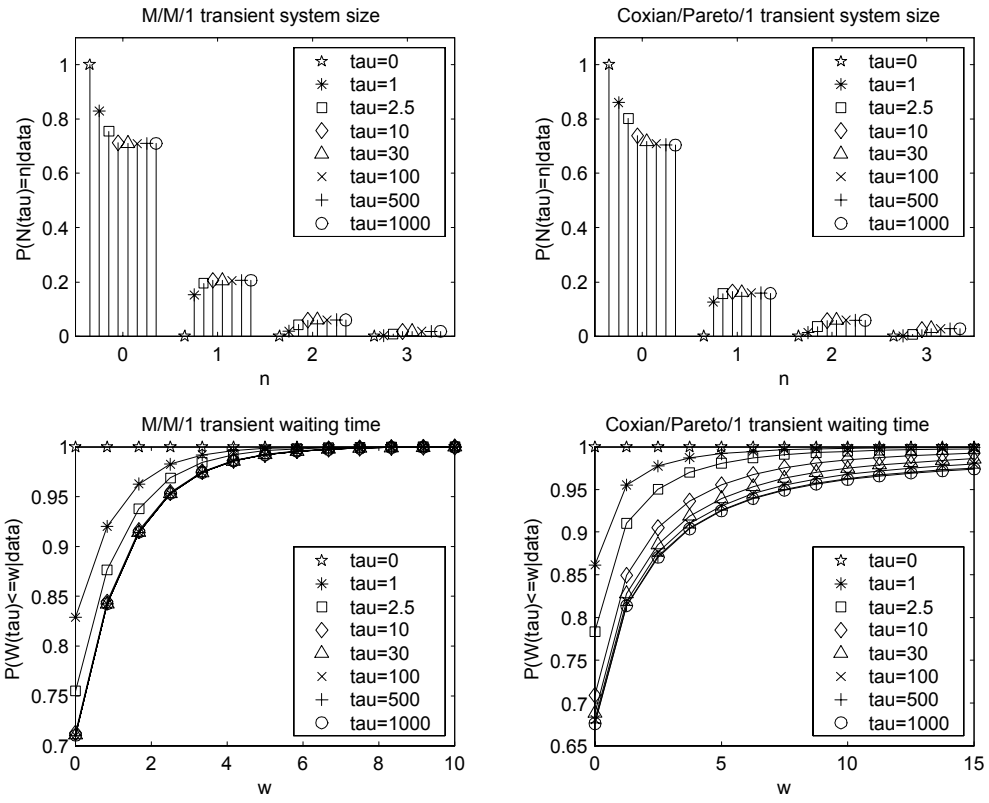

Figure 6: Transient distributions for the queue length (up) and the waiting time (down) for the two simulated systems. These approach to their stationary distributions as the time, τ, goes to infinity.

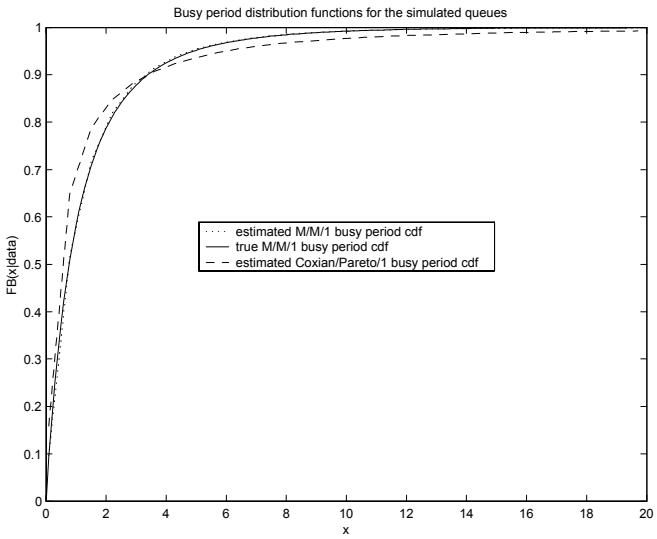

Figure 7: Estimated busy period distribution functions for the two simulated queues. Also shown is the true distribution for the M/M/1. It can hardly be distinguished from its estimation.

as given in Table 3. Note that the level of occupancy is larger for business transactions than for currency exchanges.

	$P(\rho < 1 \mid data)$	$E[\rho \mid data]$	$E[\rho < 1 \mid \rho < 1, data]$	$\hat{\rho}_{MLE}$
Foreign Currency	.99957	.59328	.59132	.58691
Business Banking	.99954	.82011	.81968	.82149

Table 3: Posterior probabilitities that the system is stable and posterior mean values for the traffic intensity for the two real bank systems.

Figure 8 illustrates the estimated transient distributions of the queue length and the waiting time, see (35) and (37), for the two real systems as a function of the time, τ . Note that the assumption of initially empty systems is true in this context. We observe that the speed of convergence is not very fast in any case. Note that when $\tau = 100$ minutes, which means that the systems have been running for more than one hour and a half, the transient distributions have not already converge to the steady-state. In particular, for

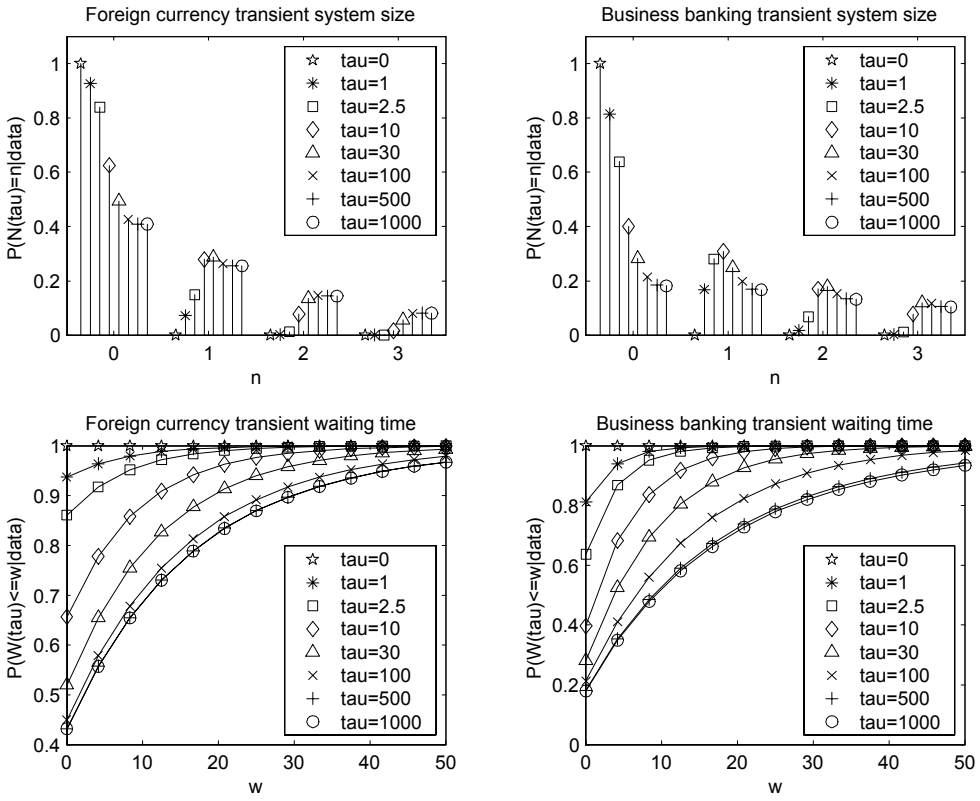

Figure 8: Transient distributions for the queue length (up) and the waiting time (down) for the two real systems. These approach to their stationary distributions as the time, τ, goes to infinity.

the business banking system, we can still appreciate differences between the distributions when τ is equal to 500 and 1000 minutes. Observe that using our approach it is possible to estimate the desired transient probability for any given instant time, τ. For example, the estimated probability that a customer who arrives at 9:00 am asking for business banking services has to wait more than 8 minutes is 0.306.

Table 4 shows some quantiles of the estimated distributions of the length of the busy period, see (38), for the two real systems. Observe that the tail of the distribution is heavier for the business bank transactions case.

Finally, we have developed a naive, ordinary $M/M/1$ analysis of the bank's queueing systems in order to investigate how our Bayesian $GI/G/1$ approach is an improvement. We have estimated both the transient

	0.25	0.50	0.65	0.80	0.90	0.95	0.97
Foreign Currency	2.3427	6.3239	10.9022	21.2713	41.1081	69.6108	96.2983
Business Banking	2.1132	4.5967	8.4113	19.4524	46.6718	95.4313	148.2902

Table 4: Quantiles of the length of the busy period distribution for the two real bank systems.

queue length and waiting time distributions and the busy period distribution using the MLE estimations of the traffic intensities given in Table 3. Results are quite different from our previous estimations. This is illustrated in Figure 9 where the estimated waiting time distributions using the two procedures are compared. We have also observed that these differences are greater for the waiting time and busy period distributions than for the queue length distribution. This should be expected because, as commented earlier, the latter is strongly dependent only on the first moments of the interarrival and service time distributions.

Acknowledgements

We wish to acknowledge the financial support provided by the project ...*Ministerio*.... and ...*Comunidad*...

References

[1] Abate, J., Choudhury, G.L., Whitt, W., 1994. Waiting-time tail probabilities in queues with long-tail service-time distributions. Queueing Systems 16, 311-338.

[2] Armero, C., Bayarri, M.J., 1994. Bayesian prediction in M/M/1 queues. Queueing Systems, 15, 401-417.

[3] Armero, C., Bayarri, M.J., 1996. Bayesian questions and Bayesian answers in queues. In: J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds., Bayesian Statistics 5, 3-23. Oxford: University Press.

[4] Armero, C., Bayarri, M.J., 1998. Dealing with uncertainties in queues and networks: a Bayesian approach. In Ghosh, S. ed., Mutivariate, Design and Sampling. Marcel Dekker, New York.

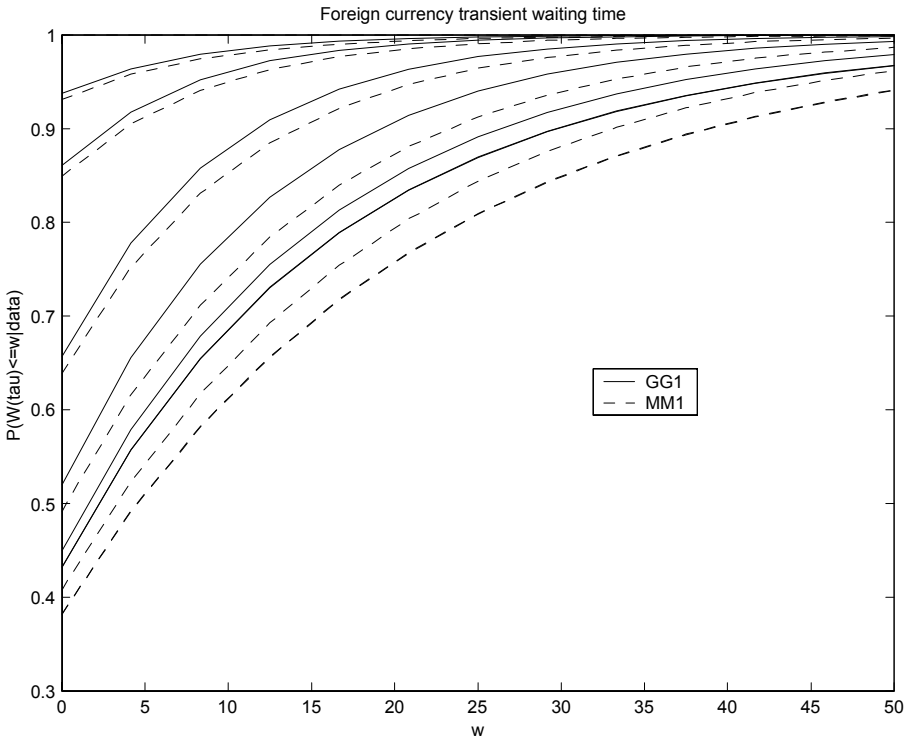

Figure 9: Comparison of the estimated transient waiting time distribution using our Bayesian GI/G/1 model and a simple M/M/1 based on the $\hat{\rho}_{MLE}$ estimator for the foreign currency exchange data.

[5] Armero, C., Conesa, D., 2000. Prediction in Markovian bulk arrival queues, Queueing Systems, 34, 327-350.

[6] Armero, C., Conesa, D., 2004. Statistical performance of a multiclass bulk production queueing system, European Journal of Operational Research, 158, 649-661.

[7] Asmussen, S. 2003. Applied probability and queues. Springer, New York.

[8] Ausín, M.C., Lillo, R.E., Ruggeri, F., Wiper, M.P., 2003. Bayesian modelling of hospital bed occupancy times using a mixed generalized Erlang distribution. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds., Bayesian Statistics 7, 443-452, Oxford University Press.

[9] Ausín, M.C., Wiper, M.P., Lillo, R.E., 2004. Bayesian estimation for the M/G/1 queue using a phase type approximation. Journal of Statistical Planning and Inference, 118, 83-101.

[10] Bertsimas, D., 1990. An analytic approach to a general class of $G/G/c$ queueing systems. Operations Research, 38, 139-155.

[11] Bertsimas, D., Nakazato, D., 1992. Transient and busy period analysis of the $G/G/1$ queue: The method of stages. Queueing Systems, 10,153-184.

[12] Borwein, P. Erdélyi, T., 1995. Polynomials and Polynomial Inequalities. Springer-Verlag, New York.

[13] Conti, P.L., 2004. Bootstrap approximations for Bayesian analysis of $Geo/G/1$ discrete-time queueing models. Journal of Statistical Planning and Inference 120, 65-84.

[14] Cumani, A., 1982. On the canonical representation of homogenous Markov processes modelling failure-time distributions. Microelectronics and reliability, 22, 583-602.

[15] Dauxois, J.-Y., 2004. Bayesian inference for linear growth birth and death processes. Journal of Statistical Planning and Inference 121, 1-19.

[16] Diebolt, J., Robert, C.P., 1994. Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society, B, 56, 363-375.

[17] Feldmann, A., Whitt, W., 1998. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. Performance Evaluation 31, 245-279.

[18] Green, P., 1995. Reversible jump MCMC computation and Bayesian model determination. Biometrika, 82, 711-732.

[19] Gross, D., Harris, C.M., 1985. Fundamentals of queueing theory. John Wiley & Sons, New York.

[20] Gruet, M.A., Philippe, A., Robert, C.P., 1999. MCMC control spreadsheets for exponential mixture estimation. Journal of Computational and Graphical Statistics, 8, 298-317.

[21] Hosono, T., 1981. Numerical inversion of Laplace transform and some applications to wave optics. Radio Science, 1015-1019.

[22] Johnson, N.L., Kotz, S., 1970. Distributions in statistics. Continuous univariate distributions. John Wiley & Sons, New York.

[23] Ralston, A., Rabinowitz, P., 1978. A first course in numerical analysis. McGraw-Hill, New York.

[24] Richardson, S., Green, P., 1997. On Bayesian analysis of mixtures with an unknown number of components. Journal of the Royal Statististical Society, B, 59, 731-792.

[25] Ríos, D., Wiper, M.P., Ruggeri, F., 1998. Bayesian analysis of $M/Er/1$ and $M/H_k/1$ queues. Queueing Systems, 30, 289-308.

[26] Robert, C.P., Mengersen, K.L., 1999. Reparameterisation Issues in Mixture Modelling and their bearing on MCMC algorithms. Computational Statistics & Data Analysis 29, 325-343.

[27] Wiper, M.P., 1998. Bayesian analysis of $Er/M/1$ and $Er/M/c$ queues, Journal of Statististical Planning and Inference, 69, 65-79.

[28] Wiper, M.P., Ríos, D., Ruggeri, F., 2001. Mixtures of gamma distributions with applications. Journal of Computational and Graphical Statistics, 10, 440-454.

4 Appendix A: Hosono Algorithm.

In order to invert these Laplace transforms, we have make use of the numerical inversion algorithm proposed by Hosono (1981) which is easy to implement and accurate in practice. This algorithm was suggested by Bertsimas and Nazakato (1992) for the case where the parameters are known. It works essentially as follows. Consider the Laplace transform of a function $f(\tau)$,

$$f^*(s) = \int_0^\infty e^{-st} f(\tau)\, d\tau.$$

The following steps allows to obtain an absolute error in the numerical inversion less than $10^{-a+1}|f(\tau)|$. In our examples, we have fixed $a = 6$.

HOSONO ALGORITHM.

For each value of τ,

1. Set $s_m = \tau^{-1}(a + i\pi(m - 0.5))$, define $F_m = (-1)^m \operatorname{Im}[f^*(s_m)]$ and find k so that,

$$\left| \sum_{r=0}^{a} \binom{a}{r} \frac{e^a}{\tau} F_{k+r} \right| < \left(\frac{2}{e^2} \right)^a.$$

2. Compute the following values,

$$C_m = 0.5^a \textstyle\sum_{r=0}^{a-m-1} \binom{a}{r}, \quad \text{for } m = 0, ..., a-1. \tag{39}$$

3. Evaluate,

$$f(\tau) \approx \frac{e^a}{\tau} \left(\sum_{m=1}^{k-1} F_m + \sum_{r=0}^{a-1} C_r F_{k+r} \right).$$

5 Appendix B. Proper posterior distribution and finite moments.

First, we show that the posterior is indeed proper. The posterior distribution is a sum over all possible sample configurations, $\mathbf{z} = (z_1, ..., z_n)$, and then, we only need to prove that the following integral is finite,

$$\int f(\lambda_1, \boldsymbol{\upsilon}) \prod_{i=1}^{n} f_{z_i}(x_i \mid \lambda_1, \boldsymbol{\upsilon}) \, d\lambda_1 d\boldsymbol{\upsilon}, \tag{40}$$

where we have integrated out the weights, $\mathbf{P}$. Assume initially that $n = 1$. Then, we observe only one observation whose density, $f_z(x \mid \lambda_1, \boldsymbol{\upsilon})$, is given in (3). Note that we consider that all rates are unequal because the set where some of the $\upsilon's$ are equal to one have measure zero. From (3), the survival distribution function is given by,

$$1 - F_z(x \mid \lambda_1, \boldsymbol{\upsilon}) = \sum_{j=1}^{z} C_j \exp\left(-\lambda_1 x \prod_{k=2}^{i} \upsilon_k\right), \tag{41}$$

where the coefficients C_j are given by,

$$C_j = \prod_{i \neq j}^{z} \frac{\prod_{k=2}^{i} \upsilon_k}{\prod_{k=2}^{i} \upsilon_k - \prod_{k=2}^{j} \upsilon_k}.$$

As for any positive continuous distribution, the distribution function, $F_z(x \mid \lambda_1, \boldsymbol{\upsilon})$, is equal to 0 for $x = 0$. Therefore, from (41), we obtain that,

$$\sum_{j=1}^{z} C_j = 1.$$

Using this property, we can now show that the integral (40) is finite for one observation,

$$\int \frac{f(\boldsymbol{\upsilon})}{\lambda_1} \sum_{j=1}^{z} C_j \lambda_1 \prod_{k=2}^{j} \upsilon_k \exp\left(-\lambda_1 x \prod_{k=2}^{j} \upsilon_k\right) d\lambda_1 d\boldsymbol{\tau} = \int f(\boldsymbol{\upsilon}) \sum_{j=1}^{z} C_j \frac{1}{x} d\boldsymbol{\upsilon} = \frac{1}{x} < \infty. \tag{42}$$

Finally, note that it is sufficient to have proved that the integral (40) is finite for $n = 1$, because now, we can define $f(\lambda_1, \boldsymbol{\upsilon} \mid x_1)$ as a new proper *prior* and consider the likelihood based on $\{x_2, ..., x_n\}$, which is regular and proper, in which case the posterior is known to be proper. Then, the integral (40) is finite for $n \geq 1$.

Now, we will prove that the mean of the predictive distribution of X is finite. Firstly, we show that the density, (3), is bounded as follows,

$$f_r(x \mid \lambda_1, \upsilon_2, ..., \upsilon_r) \leq \lambda_1, \qquad \text{for } r = 1, 2, ... \tag{43}$$

This statement can be proved by induction. For $r = 1$, we have that,

$$f_1(x \mid \lambda_1) = \lambda_1 \exp\{-\lambda_1 x\} \leq \lambda_1.$$

Now, we assume that,

$$f_{r-1}(x \mid \lambda_1, \boldsymbol{\upsilon}) \leq \lambda_1,$$

then, as f_r is the density of the sum of r exponentials, see (1), it can be expressed as the convolution of the r-th exponential density and f_{r-1},

$$\begin{aligned} f_r(x \mid \lambda_1, \boldsymbol{\upsilon}) &= \int_0^x \lambda_1 \prod_{k=2}^{r} \upsilon_k \exp\left(-\lambda_1 x \prod_{k=2}^{r} \upsilon_k\right) f_{r-1}(x-u \mid \lambda_1, \boldsymbol{\upsilon})\, du \\ &\leq \lambda_1 \left[1 - \exp\left\{-\lambda_1 x \prod_{k=2}^{r} \upsilon_k\right\}\right] \leq \lambda_1. \end{aligned}$$

The expectation of the Coxian distribution is given as the denominator (or numerator) of (17). Thus, the predictive mean of X is finite if the posterior mean of $1/\lambda_r$ exists, that is, if the following integral is finite,

$$\int \frac{f(\lambda_1, \boldsymbol{\upsilon})}{\lambda_1 \prod_{k=2}^{r} \upsilon_k} \prod_{i=1}^{n} f_{z_i}(x_i \mid \lambda_1, \boldsymbol{\upsilon})\, d\lambda_1 d\boldsymbol{\upsilon}. \tag{44}$$

Then, it is clear that if we do not observe at least two data, the predictive mean does not exist. Suppose first that we observe two data. Using (43), we have that,

$$\int \frac{f(\boldsymbol{\upsilon})}{\lambda_1^2 \prod_{k=2}^{r} \upsilon_k} f_{z_1}(x_1 \mid \lambda_1, \boldsymbol{\upsilon})\, f_{z_2}(x_1 \mid \lambda_1, \boldsymbol{\upsilon})\, d\lambda_1 d\boldsymbol{\upsilon} \leq \int \frac{f(\boldsymbol{\upsilon})}{\lambda_1 \prod_{k=2}^{r} \upsilon_k} f_{z_1}(x_1 \mid \lambda_1, \boldsymbol{\upsilon})\, d\lambda_1 d\boldsymbol{\upsilon},$$

and using the same arguments as in (42), this integral is proportional to,

$$\int \frac{f(\boldsymbol{\upsilon})}{\prod_{k=2}^{r} \upsilon_k} d\boldsymbol{\upsilon} \propto \int \prod_{k=2}^{r} \upsilon_k^{-1.1} (1-\upsilon_k)^{-0.1}\, d\boldsymbol{\upsilon} < \infty.$$

Note that this is the reason because we have chosen a $Beta(1.1, 1.1)$ for each υ_r in (7). Finally, as we know that the integral (44) is finite for $n = 2$, we can define a new proper *prior* for $(\lambda_1, \boldsymbol{\upsilon})$,

$$g(\lambda_1, \boldsymbol{\upsilon}) \propto \frac{f(\lambda_1, \boldsymbol{\upsilon})}{\lambda_1 \prod_{k=2}^{r} \upsilon_k} \prod_{i=1}^{2} f_{z_i}(x_i \mid \lambda_1, \boldsymbol{\upsilon}),$$

which is a proper density. With this prior and the likelihood based on $\{x_3, ..., x_n\}$, the posterior is proper and then, the integral (44) is finite for $n \geq 2$.